

## **Supplementary Material**

### **Contents**

1. GHMP family proteins in Thermoplasmata
2. Phylogenetic analysis of CMK (MEP enzyme)
3. Functionality of AACT and HMGS homologs in CPR and archaea
4. Synteny of genes involved in the MEP pathway (MEP genes)
5. Supplementary figure captions and figures (Figures 1-14)

### **1) GHMP family proteins in Thermoplasmata**

Thermoplasmata possesses several GHMP subfamily proteins in addition to MVK, M3K and MBD. They are annotated as Ta0344, Ta1304 and Ta0461 (Azami et al., 2014). The function of these proteins is not known. Ta0461 has a sequence similarity to MBD and is indeed included in the DPMD/MPD tree (supplementary fig. 7). In contrast, Ta0344 and Ta1304 have sequence similarity to MVK and PMVK, but only distantly. The distribution of Ta0344 homologs is limited to species from Thermoplasmata and Methanomicrobia, while Ta1304 is much more widely distributed. Outside the class Thermoplasmata, Ta1304 homologs are common in the DPANN group, but are generally rare in the archaeal and eukaryotic domains. In contrast, Ta1304 homologs are found ubiquitous in both the bacterial domain and the superphylum CPR. However, MVA gene-bearing species in these domains rarely possess Ta1304 homologs. Thus, the distribution of Ta1304 and other GHMP proteins seems reciprocal. A phylogenetic analysis of Ta1304 did not reflect much similarity to the species tree (data not shown). Therefore, Ta1304 is inferred to have had a complex evolutionary history different from other GHMP family proteins and experienced substantial gene exchanges among different bacterial phyla and domains of life.

## 2) Phylogenetic analysis of CMK (MEP enzyme)

A phylogenetic tree of the sole GHMP homolog (CMK) found in the MEP pathway was reconstructed as a reference to multiple GHMP subfamilies found in the MVA pathway (supplementary fig. 14). CMK is universally distributed in the bacterial domain. In our current analysis, major well-established bacterial superphyla were chosen for the tree construction; Terrabacteria, FCB group, PVC group and Proteobacteria ( $\alpha$ -,  $\beta$ -, and  $\gamma$ -proteobacteria). The tree topology within individual bacterial phyla and superphyla is in good agreement with the suggested species relationship (Hug et al., 2016). This observation indicates that the CMK gene is mostly vertically transmitted in the bacterial domain and the influence of HGT is small. The relationship between each superphylum also broadly matches the species relationship, except for the FCB group. Therefore, it is likely that CMK existed in the bacterial common ancestor. CMK homologs found in archaea (only DPANN group) and CPR also form distinct clades, respectively. Hence, the common ancestor of the DPANN group and the CPR superphylum may also have possessed CMK. Considering the possibility that the DPANN group and CPR may represent the deepest branch of the archaeal and bacterial domains, respectively, LUCA might have possessed CMK and thus potentially the MEP pathway. However, the distribution of CMK and other MEP enzymes is generally very limited in both CPR and the archaeal domain. Thus, the presence of MEP pathway genes in LUCA is not conclusive. In our current study, it is more likely that the CMK gene in archaea and CPR was obtained via ancient HGT from FCB group bacteria (supplementary fig. 14). A more comprehensive analysis including all bacterial phyla for all seven MEP enzymes may provide further insight.

There are two eukaryotic lineages in the CMK phylogeny (supplementary fig. 14). Nearly all eukaryotic CMK sequences cluster with those of chlamydiae, whereas only one eukaryotic sequence clusters near cyanobacteria. This is contrary to the general assumption that MEP genes originate from an endosymbiotic cyanobacterium in an early eukaryotic cell that led to the modern plastid in photosynthetic eukaryotes (Lichtenthaler, 1999). However, the contribution of chlamydia genes to the eukaryotic plastid has been suggested (Horn et al., 2004; Moustafa et al., 2008). Therefore, the presence of eukaryotic CMK next to chlamydia homologs may be explained by possible ancestral gene exchanges between chlamydia and the ancestral endosymbiotic cyanobacteria or between chlamydia and an early plastid-bearing eukaryote.

### 3) Functionality of AACT and HMGS homologs in CPR and archaea

The biochemical function of archaeal AACT homologs is largely unknown, except for one halobacterial enzyme that was functionally characterized (Liu et al., 2002). A phylogenetic study has identified a conserved motif of Cys-His diad (CH motif) within AACT sequences while a Cys-His-Asn triad (CHN motif) is found within HMGS sequences (Jiang et al., 2008). In our current study, all AACT homologs in CPR and the majority of archaeal AACT homologs that are encoded by the genes in synteny with HMGS genes are confirmed to possess the CH motif (supplementary fig. 3). However, there are some notable exceptions that lack this critical motif, including proteins from several species of Thermoplasmata, Crenarchaeota (TACK group) and Chloroflexi. These proteins lack one or both residues of the motif and are characterized by a relatively high amount of sequence divergence compared to nearby homologs containing the CH motif. Some of these species possess additional archaeal AACT paralogs that retain the CH motif but these paralogs are not in synteny with HMGS (Other AACT homologs; supplementary fig. 3). One possibility is that these paralogs functionally compensate for the lack of the CH motif from the AACT homologs in synteny with HMGS. Another possibility is that thiolase II compensates for the lack of the CH motif in AACT. Most of the species mentioned above have proteins that are homologous to the thiolase II enzymes from bacteria and eukaryotes. The thiolase II homologs in these archaea and chloroflexi form two distinct groups in the thiolase II tree (supplementary fig. 4). Therefore, it is possible that the function of the original archaeal AACT was replaced by thiolase II in these species. Surprisingly, halobacterial HMGS genes in synteny with AACT genes do not contain the conserved CHN motif (Halobacteria I; fig. 3). Yet, more-distant homologs outside the archaeal clade retain this motif (Halobacteria II). The precise functions of AACT and HMGS homologs missing these critical motifs require biochemical characterization in future studies.

#### 4) Synteny of MEP genes

MEP genes are distributed universally in bacteria and a small number of CPR & DPANN species from several discrete phyla. In bacteria, MEP genes never form a single gene cluster that includes all MEP genes. Instead, some MEP genes form gene clusters with only one or two other MEP genes and/or isoprenyl diphosphate synthase (IPPS) genes. There are three distinct gene clusters commonly observed in bacteria. They are named Cluster 1-3 for convenience in this text. Cluster 1 comprises the first MEP gene, DXS, and the *trans*-IPPS gene (see supplementary fig. 1 for MEP gene abbreviations). Cluster 2 contains the third and the fifth genes, MCT and MDS. Cluster 3 includes the second gene, DXR, and the *cis*-IPPS gene. In Terrabacteria, Cluster 3 also commonly includes the sixth gene, HDS (Firmicutes, Actinobacteria other than the class Actinobacteria, and Chloroflexi). The fourth and the seventh genes, CMK and HDR, rarely form a gene cluster, except for a few phyla (e.g. Deinococcus-Thermus). Although there are small variations among bacterial phyla, these three clusters are widely distributed in bacteria, particularly in Terrabacteria and Proteobacteria. However, the synteny of MEP genes is not universal. Gene clusters are absent or rare in Cyanobacteria, FCB group and PVC group.

In contrast, MEP genes in CPR and the DPANN group (a deep-branching archaeal group) mostly form a single gene cluster containing all or most MEP genes. The cluster further commonly contains the *cis*-IPPS gene in CPR, as is the case for the MVA gene cluster in CPR. MEP genes in the DPANN group do not have a syntenic relationship to *trans/cis*-IPPS genes.

## References

- Azami Y, Hattori A, Nishimura H, Kawaide H, Yoshimura T, Hemmi H (2014) (R)-Mevalonate 3-Phosphate Is an Intermediate of the Mevalonate Pathway in *Thermoplasma acidophilum*. *The Journal of Biological Chemistry*, **289**, 15957-15967.
- Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D, Rattei T, Mewes H-W, Wagner M (2004) Illuminating the Evolutionary History of Chlamydiae. *Science*, **304**, 728-730.
- Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF (2016) A new view of the tree of life. *Nature Microbiology*, **1**, 16048.
- Lichtenthaler HK (1999) The 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, **50**, 47-65.
- Moustafa A, Reyes-Prieto A, Bhattacharya D (2008) Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. *PLoS ONE*, **3**.

## Supplementary figure captions

**Figure 1.** MEP pathway. Abbreviations for enzymes: DXS, deoxyxylulose 5-phosphate synthase; DXR, deoxyxylulose 5-phosphate reductoisomerase; CMT, 2-C-methylerythritol 4-phosphate cytidyl transferase; CMK, 4-(cytidine 5'-diphospho)-2-C-methylerythritol kinase; MCS, 2-C-methylerythritol 2,4-cyclodiphosphate synthase; HDS, 4-Hydroxy-3-methylbut-2-enyl diphosphate synthase; HDR, 4-Hydroxy-3-methylbut-2-enyl diphosphate reductase.

**Figure 2.** Distribution of the MVA pathway in the bacterial domain. The cladogram of bacterial phyla is modified from (Hug et al. 2016) to indicate the general branching order of each bacterial phylum. Bacterial phyla in which at least one species harbors three or more MVA enzymes are indicated by a box for each enzyme. The grey box indicates that the corresponding MVA genes are found in only a couple of discrete species or genera within a phylum. It is noted that Tectomicrobia, Caldithrichaeota, Zixibacteria and KSB1 have only one species containing MVA genes within each phylum, but the available genomes are less than 20 for each of these four phyla. Hence the distribution of MVA genes in these phyla is less clear than for other phyla.

**Figure 3.** Bayesian phylogenetic tree of archaeal AACT with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. The left bar indicates the presence/absence of the conserved CH motif (red/blue color). The right bar indicates the presence/absence of a syntenic HMGS gene (red/blue color). Abbreviations used in common in all supplementary figures: Bac = Bacteria; Arc = Archaea; Euk = Eukaryotes; CPR = Candidate Phyla Radiation; Fm = Firmicutes; Tn = Tenericutes; Cy = Cyanobacteria; Clf = Chloroflexi; Ac = Actinobacteria; Fs = Fusobacteria; Zx = Zixibacteria; Cla = Clostrimonetes; Bln = Balneolaeota; Cld = Caldithrichaeota (Caldithrix); Pl = Planctomycetes; V = Verrucomicrobia; Chl = Chlamydiae; Le = Lentisphaerae; Tc = Tectomicrobia; Dp = Dependentiae (TM6); aP =  $\alpha$ -Proteobacteria; bP =  $\beta$ -Proteobacteria; gP =  $\gamma$ -Proteobacteria; dP =  $\delta$ -Proteobacteria; Sp = Spirochaetes; Bc = Bacteroidetes; Clb = Chlorobi; Ig = Ignavibacteria; Fb = Fibrobacteres; Pc = Parcubacteria; Mc = Microgenomates; DPN = DPANN group archaea; Asg = Asgard group archaea; Ery = Euryarchaeota; TK = TACK group archaea; Fg = Fungi; SAR = SAR group; Arpl = Archaeplastida; Met = Metazoa.

**Figure 4.** Bayesian phylogenetic tree of bacterial/eukaryotic-type AACT (thiolase II) with sequence annotations. Sequences from species that lack the MVA pathway are not included in the dataset. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. Abbreviations, T1 =

thiolase I; T2 = thiolase II, Cyt = cytosolic; Oth = Other cellular compartments. For other abbreviations, see supplementary fig. 3.

**Figure 5.** Bayesian phylogenetic tree of HMGS with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. The left bar indicates the presence/absence of the conserved CHN motif (red/blue color). The right bar indicates the presence/absence of a syntenic archaeal AACT gene (red/blue color). For abbreviations, see supplementary fig. 3.

**Figure 6.** Bayesian phylogenetic tree of HMGR-I with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 7.** Bayesian phylogenetic tree of HMGR-II with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 8.** Bayesian phylogenetic tree of MVK with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 9.** Bayesian phylogenetic tree of PMVK with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 10.** Bayesian phylogenetic tree of DPMD/MPD homologs with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 11.** Bayesian phylogenetic tree of IDI-I with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 12.** Bayesian phylogenetic tree of IDI-II with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 13.** Bayesian phylogenetic tree of IPK with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

**Figure 14.** Bayesian phylogenetic tree of CMK with sequence annotations. Numbers at nodes are posterior probabilities. Nodes with the posterior probabilities of less than 0.5 are collapsed. Scale bar represents 0.5 amino acid replacements per site per unit evolutionary time. For abbreviations, see supplementary fig. 3.

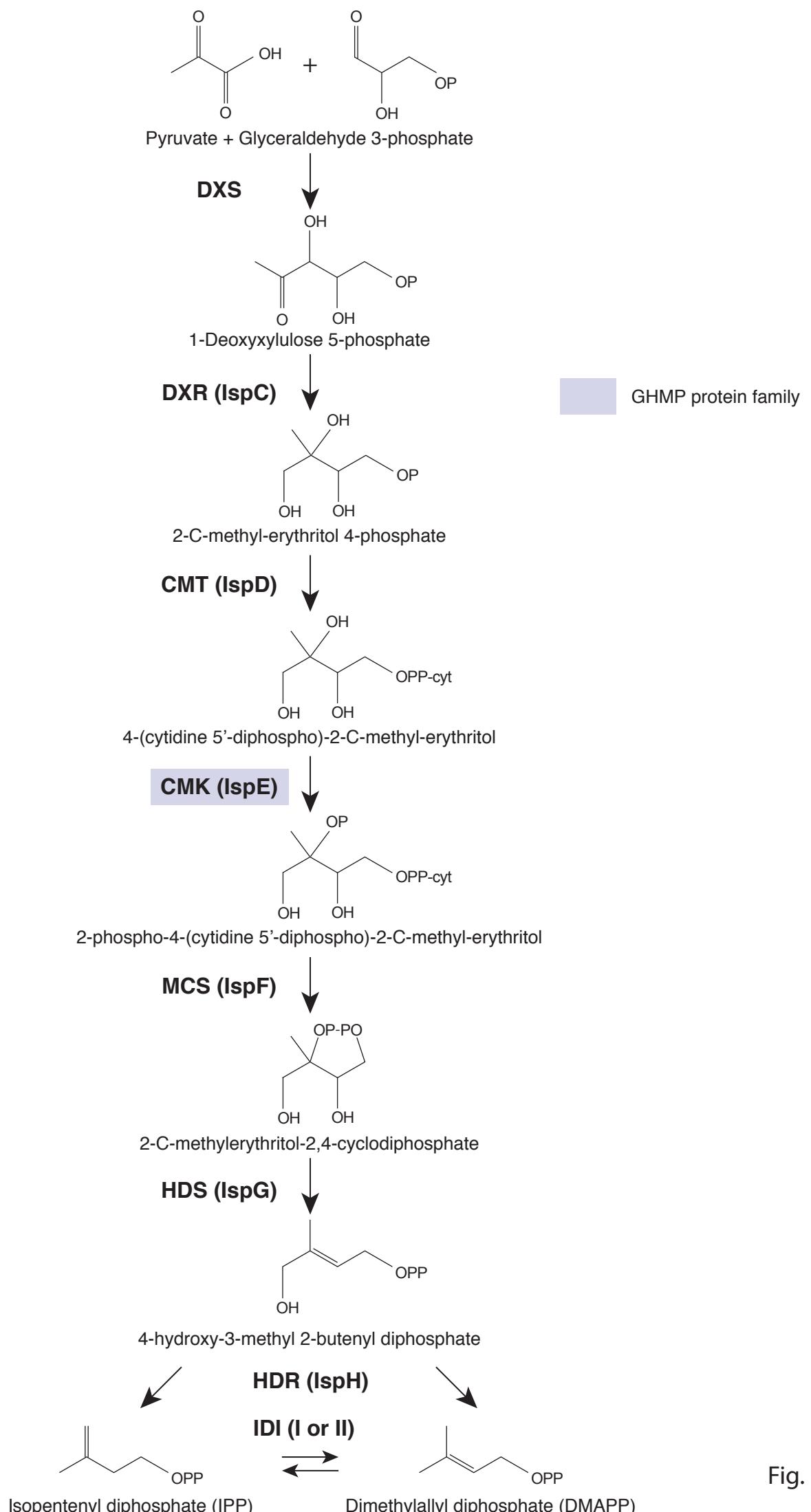


Fig. S1

## MVA pathway

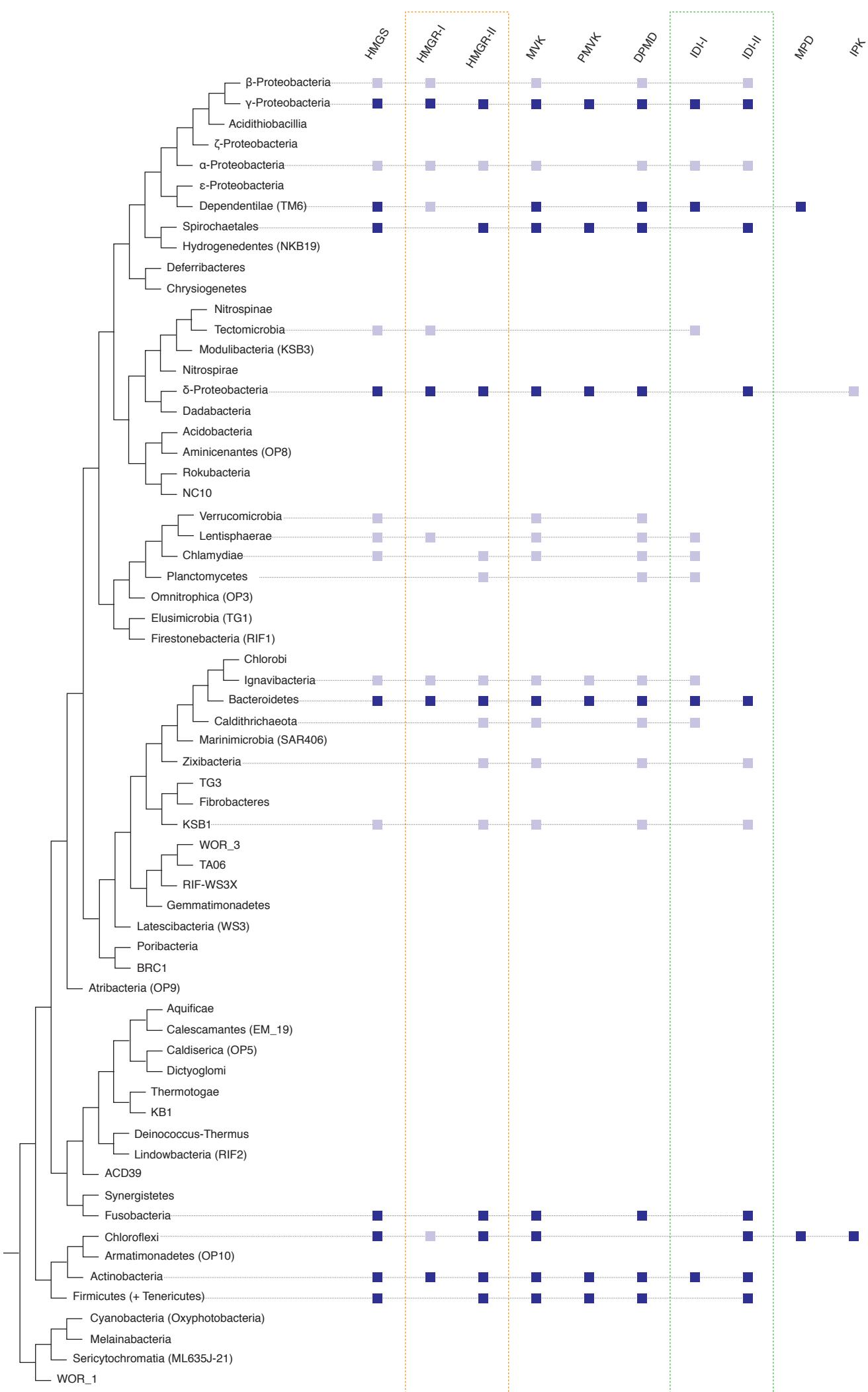


Fig. S2

# Archaeal AACT

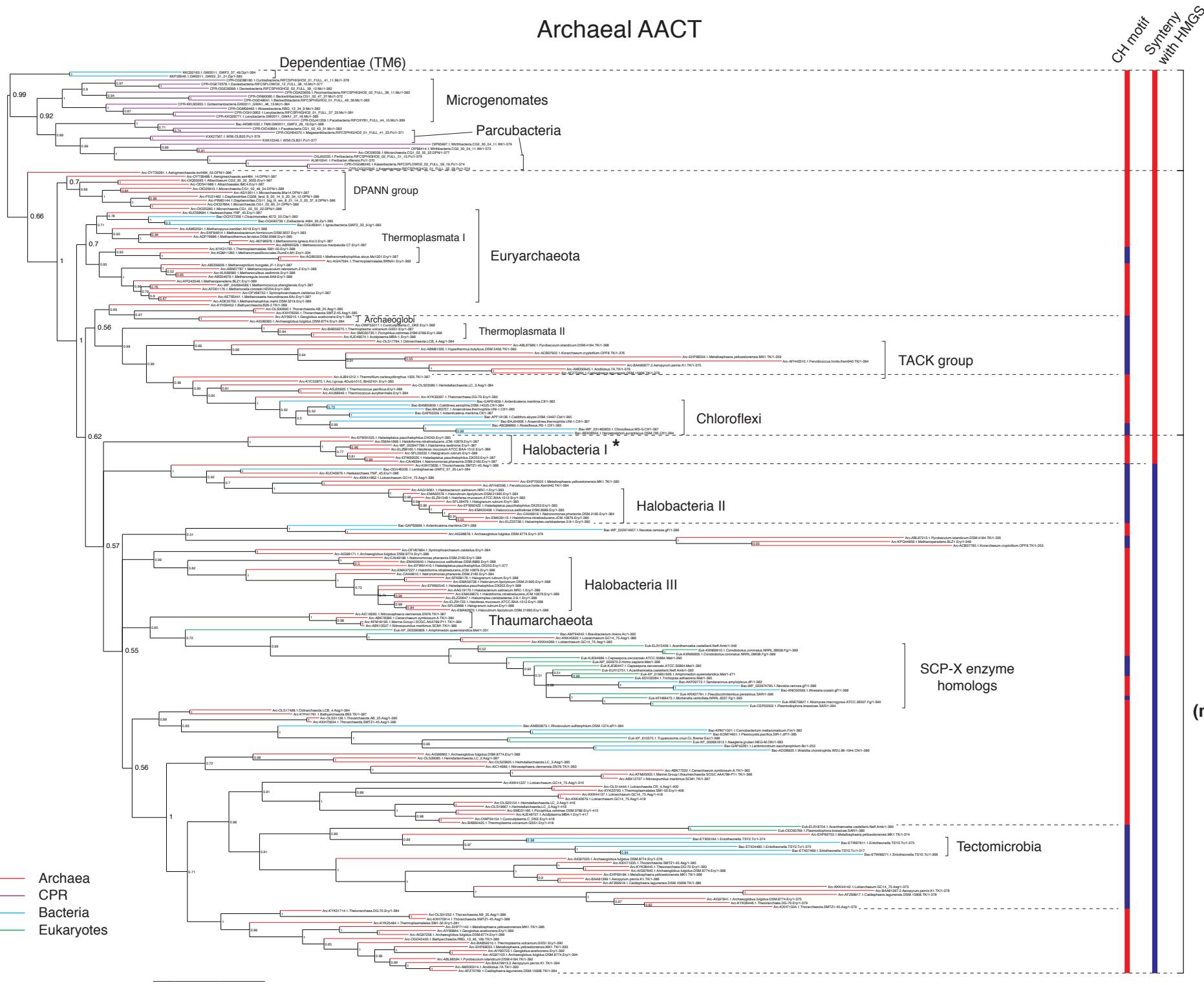


Fig. S3

## Thiolase II

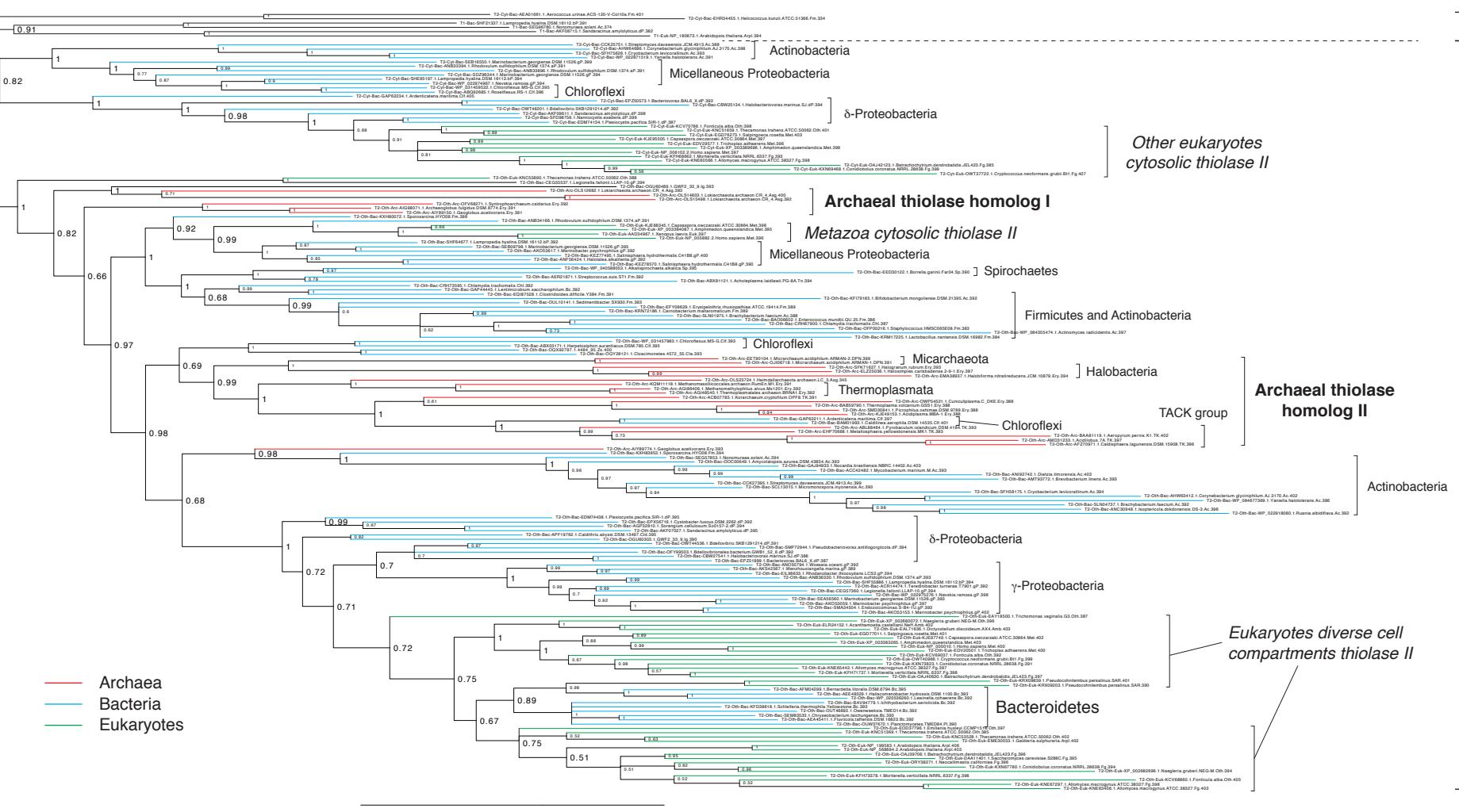
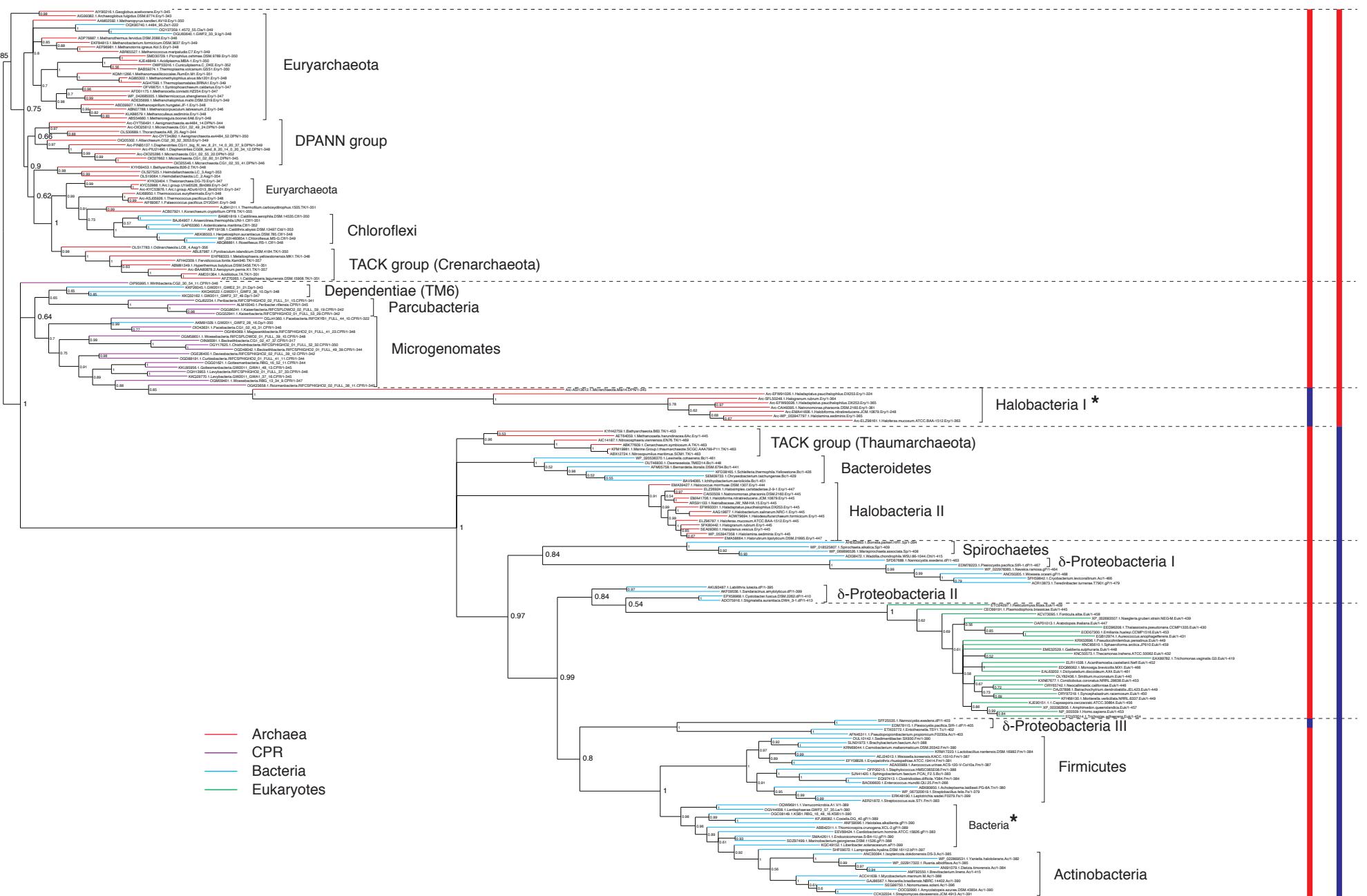


Fig. S4

# HMGS

CHN motif  
Synteny  
with ACT



# HMGR-I

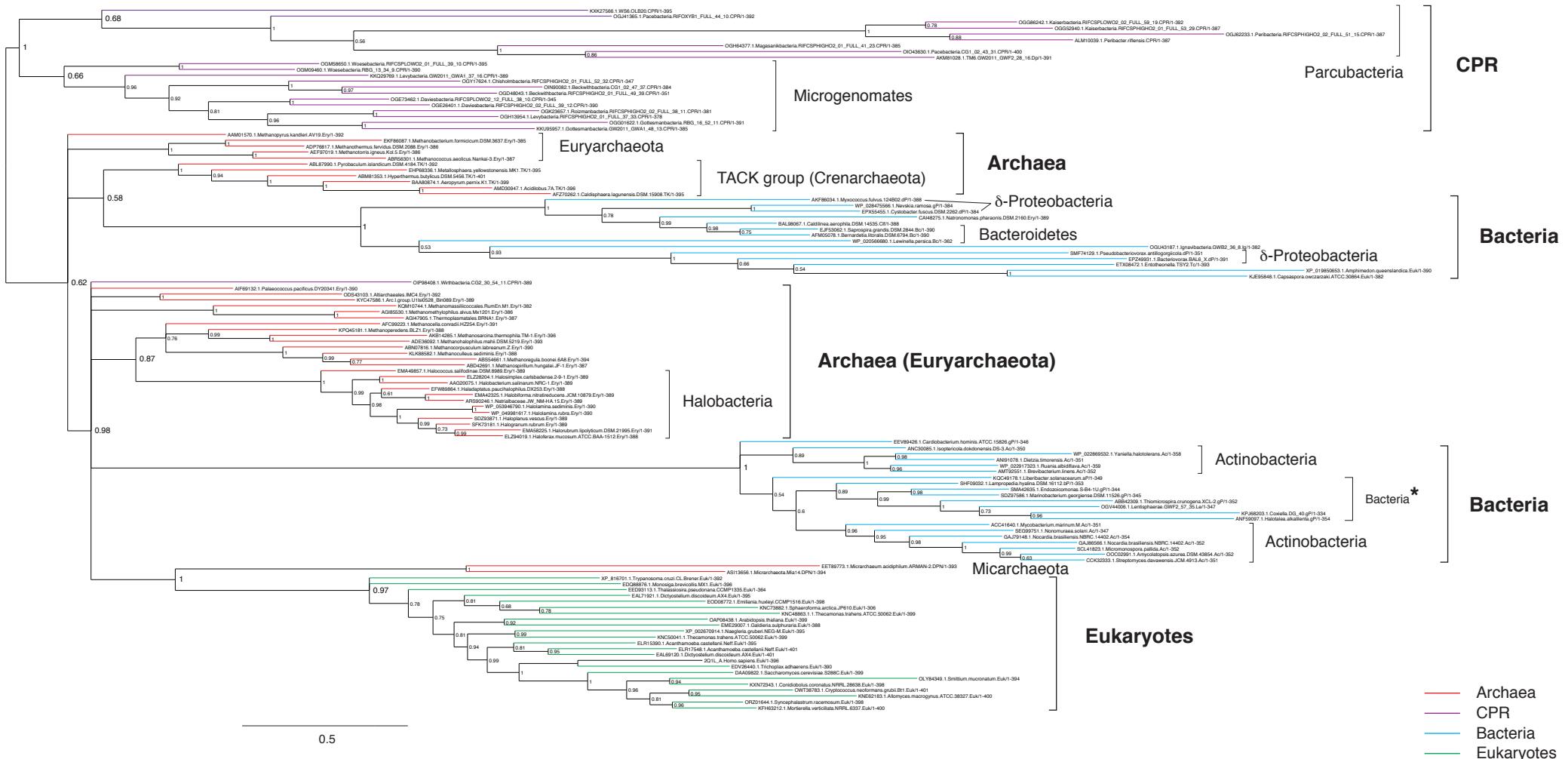


Fig. S6

HMGR-II

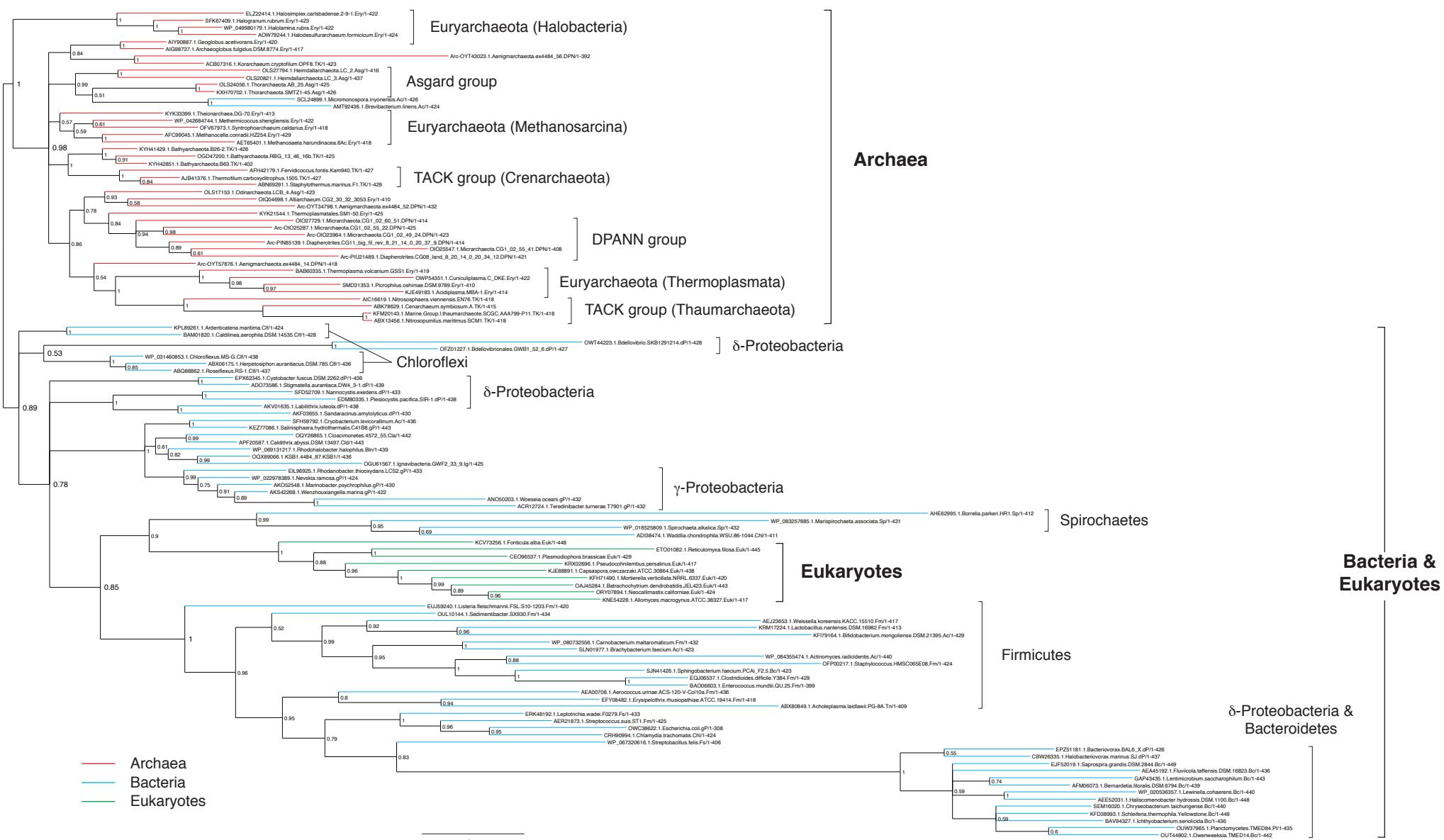


Fig. S7

# MVK

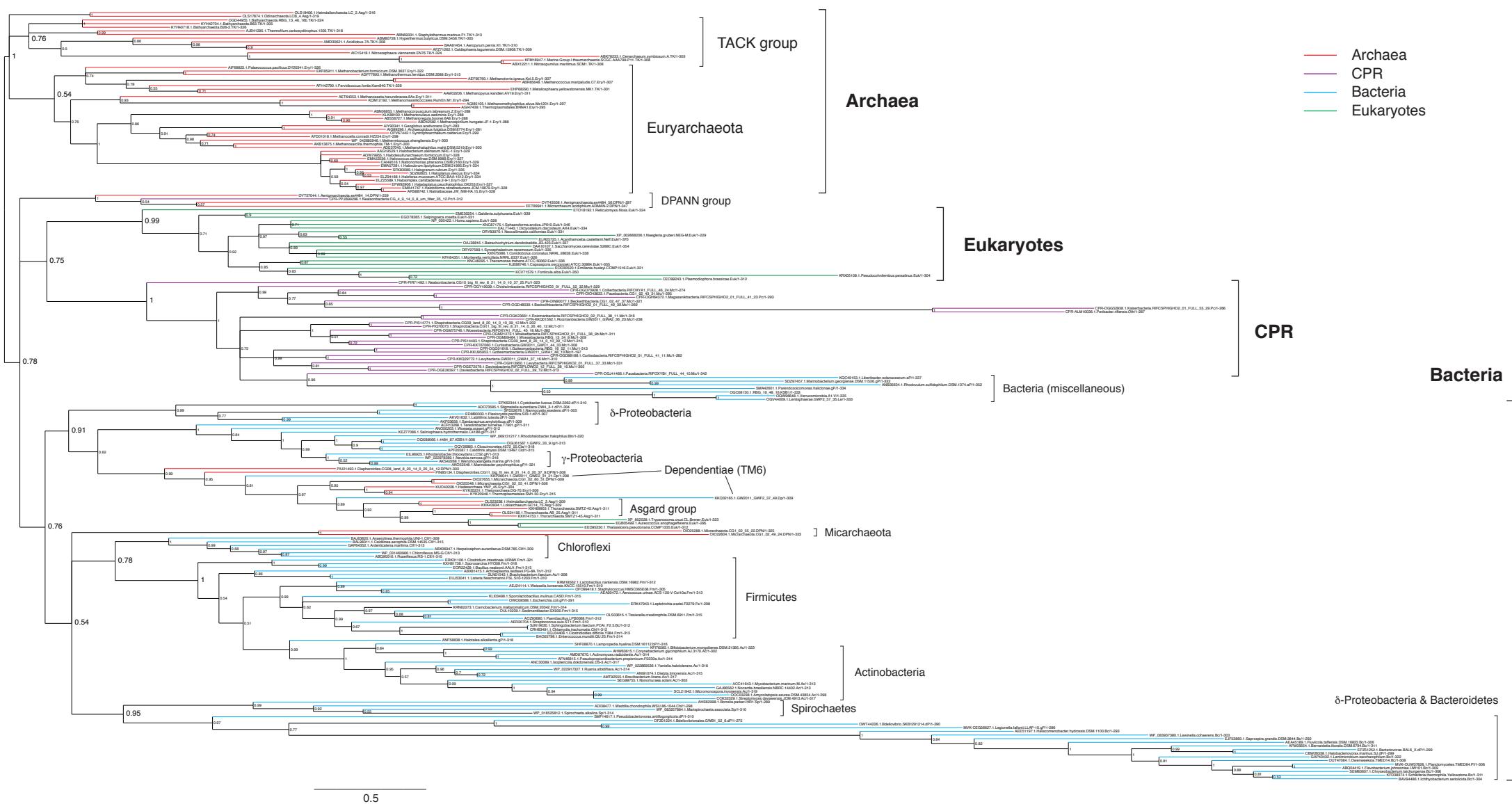


Fig. S8

PMVK

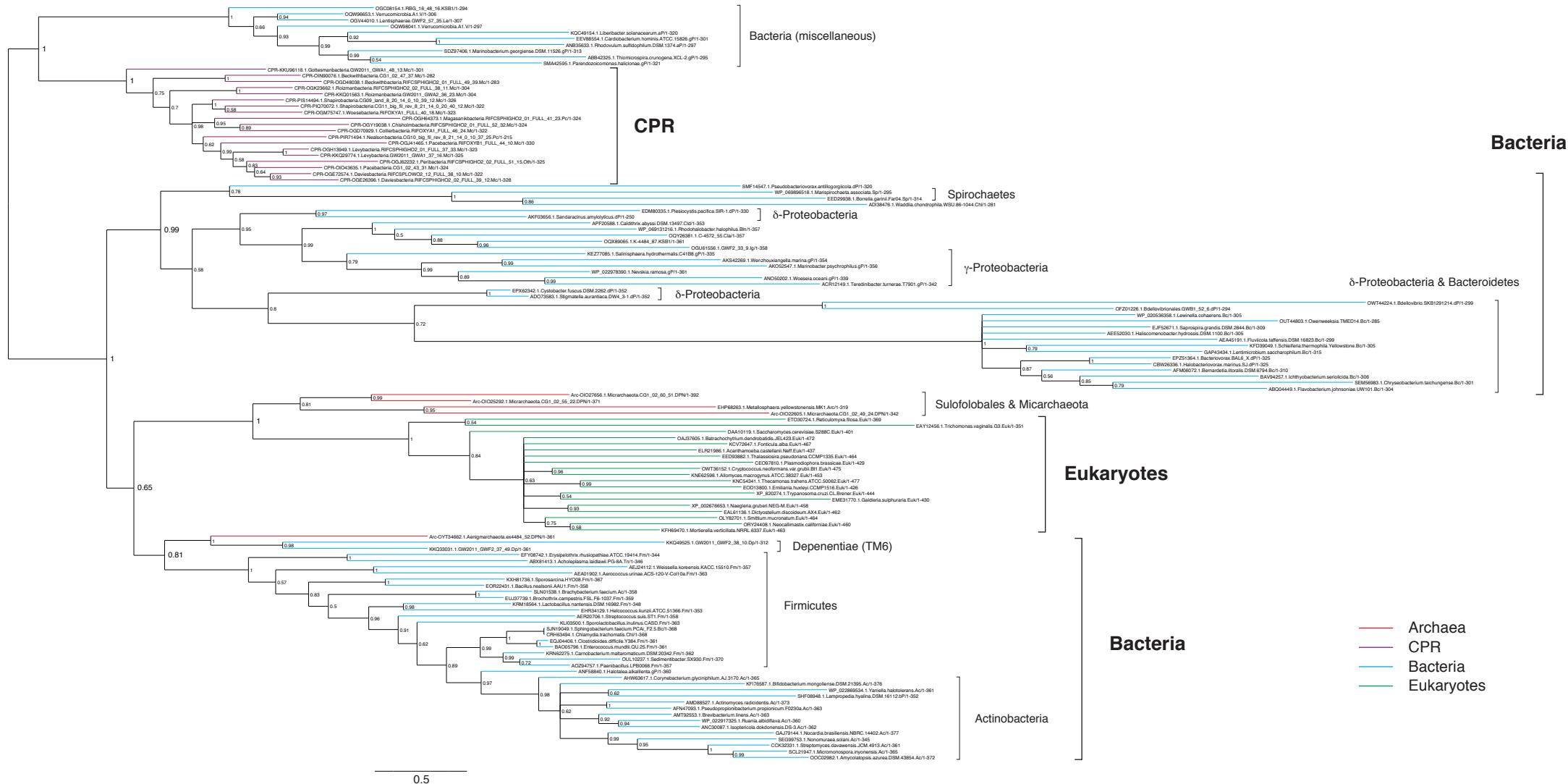


Fig. S9

# DPMD/MPD/M3K/MBD

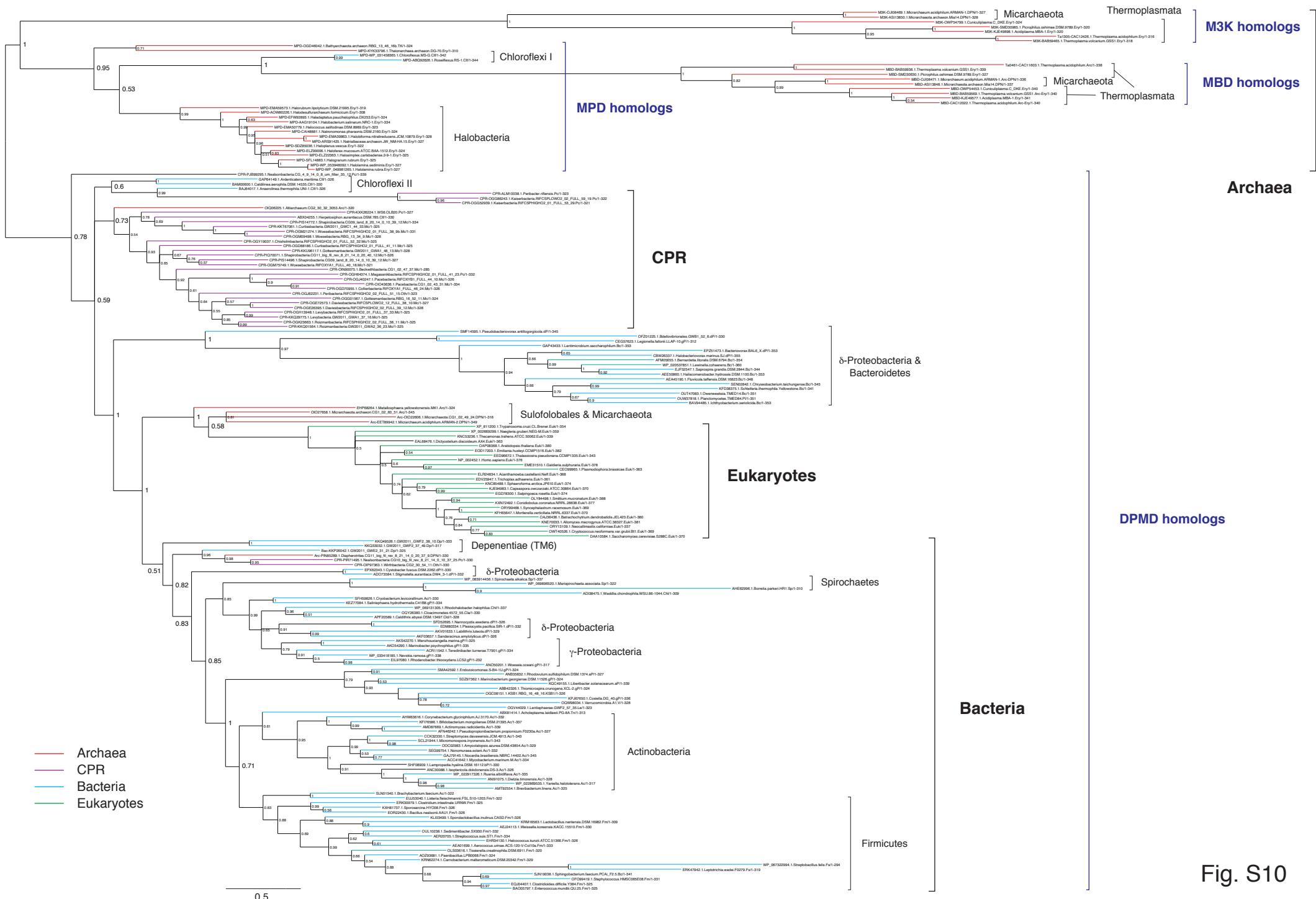


Fig. S10

# IDI-I

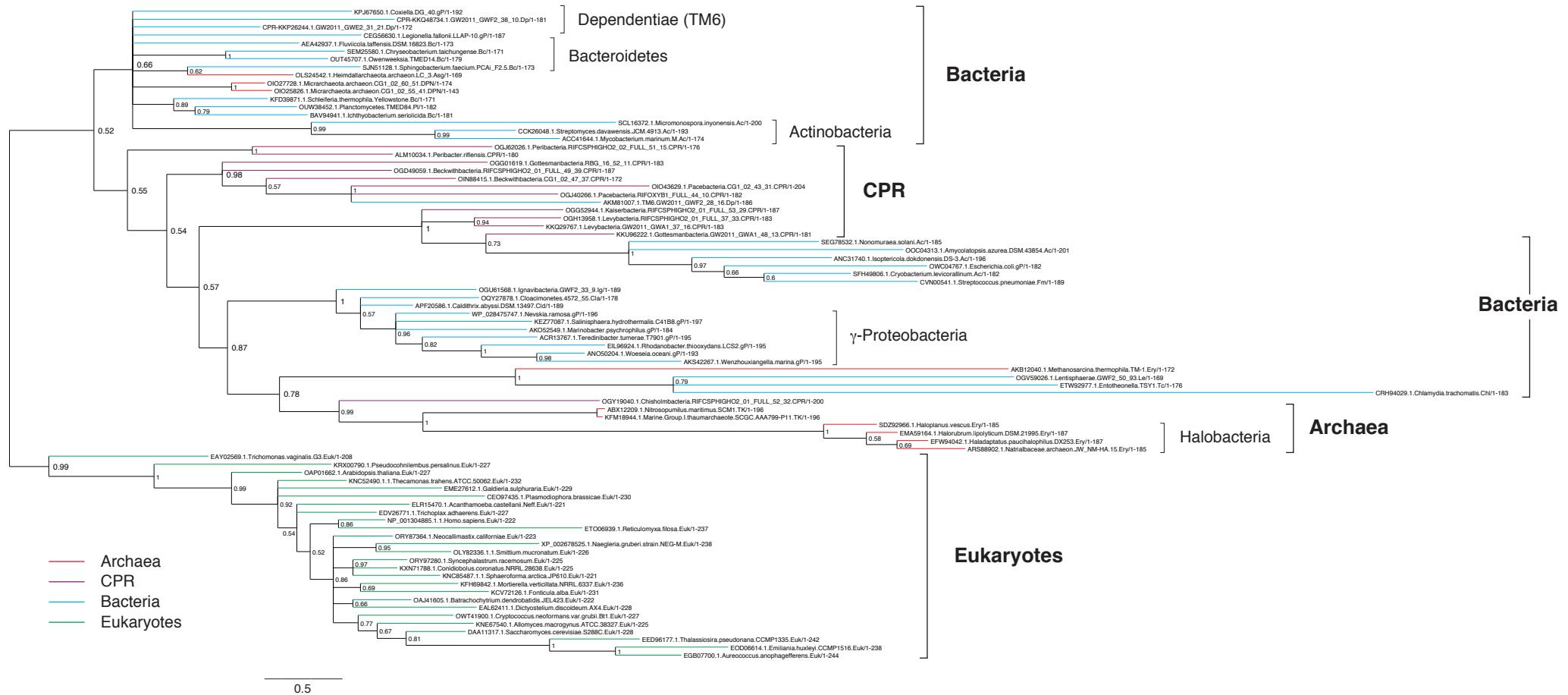


Fig. S11

IDI-II

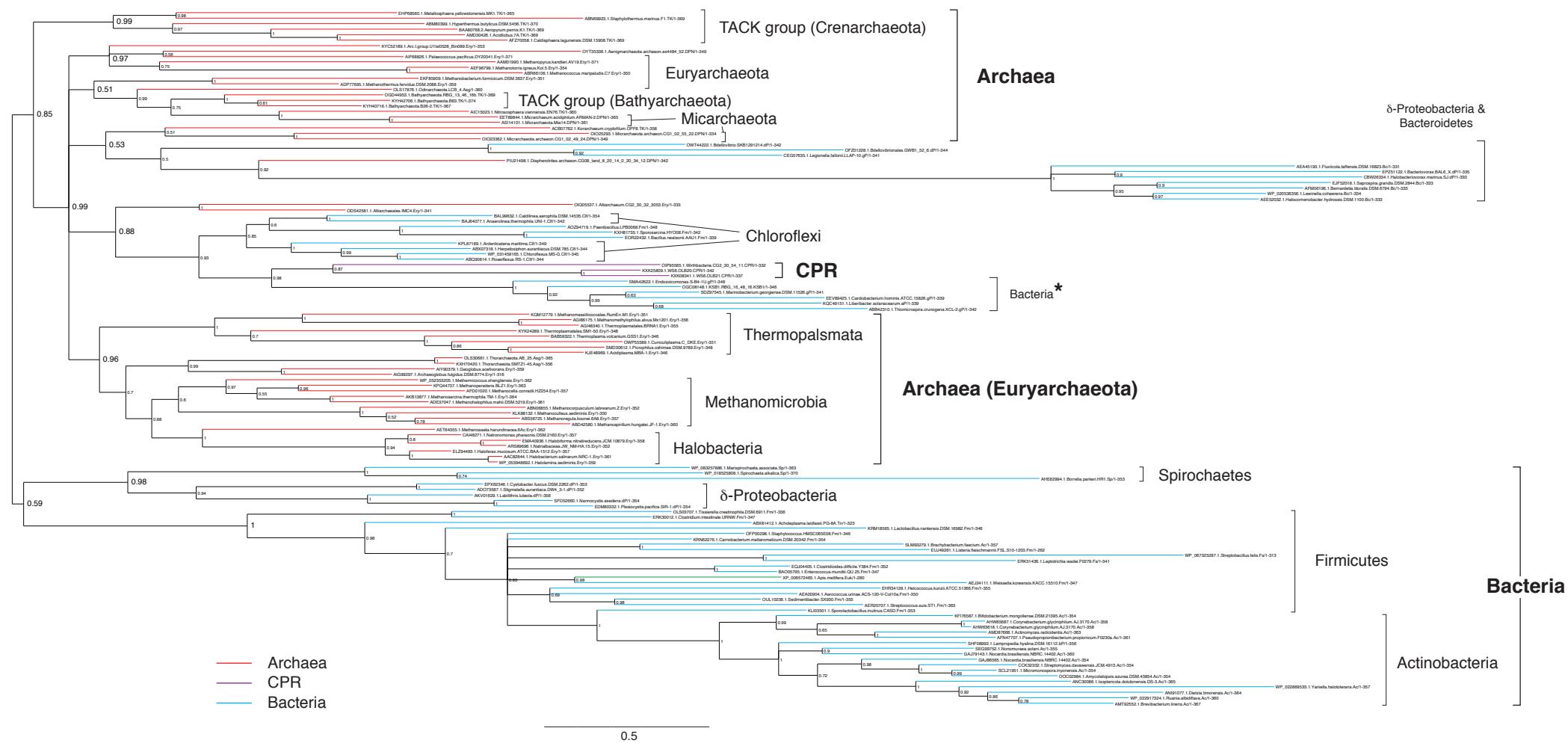


Fig. S12

IPK

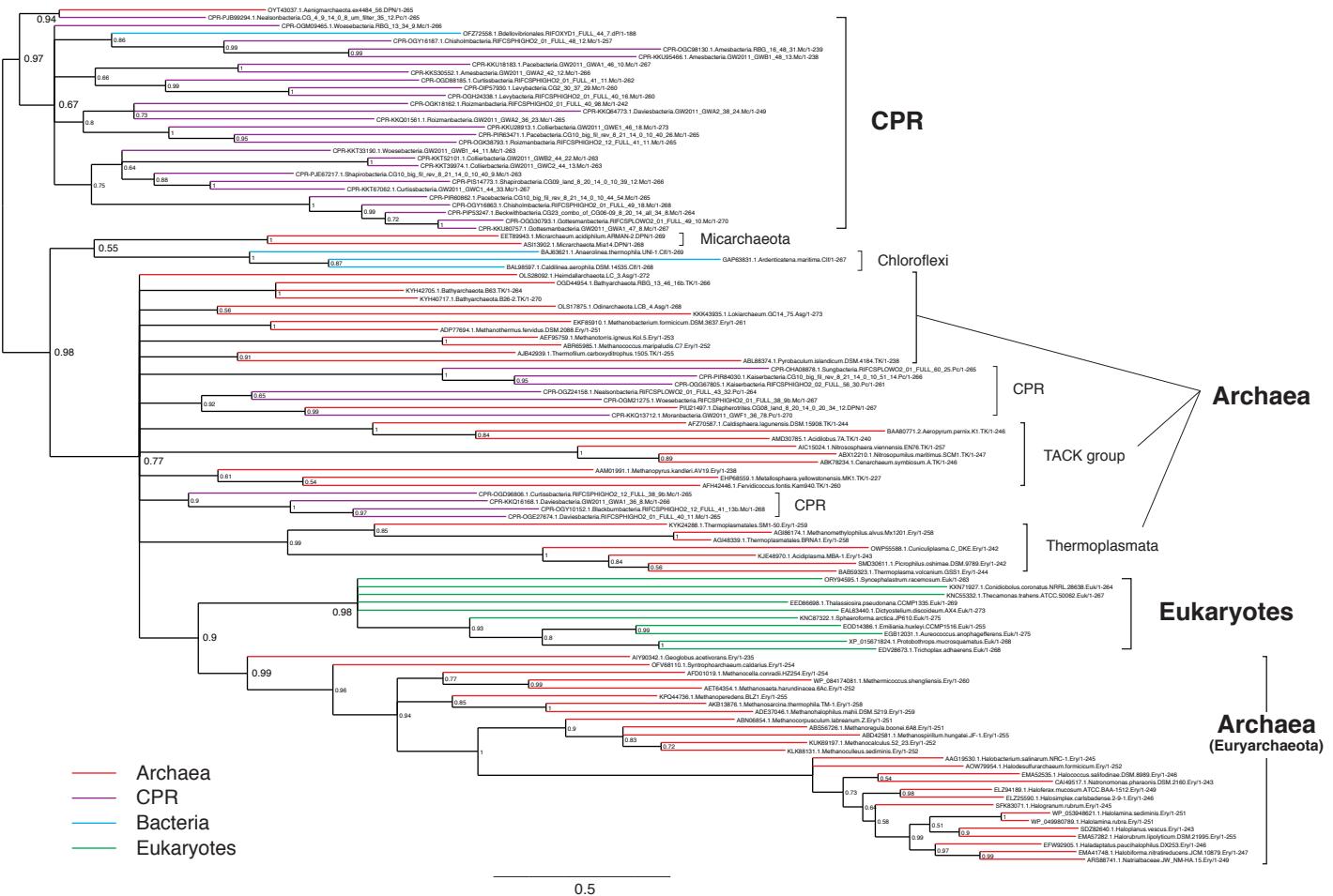


Fig. S13

# CMK

Bacterial superphylum

Terrabacteria

CPR

FCB group

DPANN group (Archaea)

CPR

Terrabacteria

PVC group

CPR

Proteobacteria

- Archaea
- CPR
- Bacteria
- Eukaryotes

0.5

Fig. S14